

American Journal of Evaluation

<http://aje.sagepub.com>

Prose and Cons about Goal-Free Evaluation

Michael Scriven

American Journal of Evaluation 1991; 12; 55

DOI: 10.1177/109821409101200108

The online version of this article can be found at:
<http://aje.sagepub.com/cgi/content/abstract/12/1/55>

Published by:



<http://www.sagepublications.com>

On behalf of:

[American Evaluation Association](#)

Additional services and information for *American Journal of Evaluation* can be found at:

Email Alerts: <http://aje.sagepub.com/cgi/alerts>

Subscriptions: <http://aje.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

TRACES are what evaluators left behind—discoveries, records, tracks—which made marks on the profession of program evaluation. Published here are excerpts from our past (e.g., articles, passages from books, speeches) that show where we have been or affected where we are g(r)owing. Suggestions for inclusion should be sent to the Editor, along with rationale for their import. Photocopies of the original printed versions are preferred with full bibliographic information. Copying rights will be secured by the Editor.

Editor's Note: The two articles that follow are very early copies of Michael Scriven's and Robert Stake's writings on goal-free and responsive evaluation, respectively. Many thanks to the UCLA Center for the Study of Evaluation for permission to copy Scriven's paper from the 1972, Volume 3, Number 4, of *Evaluation Comments* and to Kluwer Academic Publishers for permission to reprint the excerpt from Chapter 17 of *Evaluation Models*, a 1983 book edited by George F. Madaus, Michael Scriven, and Daniel L. Stufflebeam.

Prose and Cons about Goal-Free Evaluation

MICHAEL SCRIVEN

INTRODUCTION

In the winter of 1970-71, the National Center for Educational Communications of USOE asked ETS to evaluate the disseminable products of the regional labs and R&D centers. The reward for success was to be substantial grants to assist dissemination. ETS set up an external committee to do the evaluation, under the chairmanship of David Krathwohl, and provided very extensive and excellent staff support for what had to be a rather rapid review. In order to standardize the practice as well as the products of the committee (on which I served) I began to develop a standard form to serve as a check list for us and, when filled out, as a summary for ETS and NCEC. There were originally about 70 entries in what became known as the Product

Michael Scriven • Director, Evaluation Institute, Pacific Graduate School of Psychology, Palo Alto, CA.

Evaluation Practice, Vol. 12, No. 1, 1991, pp. 55-76.
ISSN: 0191-8036

Copyright © 1991 by JAI Press, Inc.
All rights of reproduction in any form reserved.

Evaluation Pool, and they ranged from toys for pre-schoolers through publications on teacher training and bilingual curricula, to vast new systems for managing schools. On these, we had varying amounts of data about field trials, mostly very thin, we had the write-ups by the producing staff and other observers, and we had the products themselves. Other input was the list of current USOE priorities in education.

It seemed very natural to start off the evaluation form with a rating of goals of the project and to go on with the rating of the effectiveness in meeting them, costs, etc. By the sixth draft of the form, another item had become very prominent, namely side-effects. Naturally, these had also to be rated, and in one case a product finished up in the Top Ten in spite of zero results with respect to its intended outcomes because it did so well on an unanticipated effect.

INTENDED AND UNINTENDED EFFECTS—WHY DISTINGUISH?

Reflecting on this experience later, I became increasingly uneasy about the separation of goals and side-effects. After all, we weren't there to evaluate goals as such—that would be an important part of an evaluation of a *proposal*, but not (I began to think) of a *product*. All that should be concerning us, surely, was determining exactly what effects this product had (or most likely had), and evaluating those, whether or not they were intended.

In fact, it was obvious that the rhetoric of the original proposal which had led to a particular product was frequently put forward as if it somehow constituted supporting evidence for the excellence of the product. This rhetoric was often couched in terms of the "in" phrases of five-year-old educational fads, sometimes given a swift updating with references to the current jargons or lists of educational priorities. That is, the rhetoric of intent was being used as a substitute for evidence of success. Was it affecting us? It would be hard to prove it didn't. And it contributed nothing, since we were not supposed to be rewarding good intentions.

Furthermore, the whole language of "side-effect" or "secondary effect" or even "unanticipated effect" (the terms were then used as approximate synonyms) tended to be a put-down of what might well be the crucial achievement, especially in terms of new priorities. Worse, it tended to make one look less hard for such effects in the data and to demand less evidence about them—which is extremely unsatisfactory with respect to the many potentially very harmful side-effects that have turned up over the years.

It seemed to me, in short, that consideration and evaluation of goals was an unnecessary but also a possibly contaminating step. I began to work on an alternative approach—simply, the evaluation of *actual* effects against (typically) a profile of *demonstrated* needs in this region of education. (This is close to what Consumers' Union actually does.) I call this Goal-Free Evaluation (GFE).

GOAL-FREE FORMATIVE EVALUATION

At first, it seemed that the proper place for goal-free evaluation (GFE) was in the summative role, like the NCEC activity. In the formative situation, the evaluator's principal task must surely be telling the producer whether the project's goals were being met.

But the matter is not so simple. A crucial function of good formative evaluation is to give the producer a preview of the summative evaluation. Of course, a producer has made the bet that if the goals of the project are achieved, the summative evaluation will be or should be favorable. But one can scarcely guarantee the non-occurrence of undesirable side-effects—and one should not overlook the possibility of desirable ones that can be cultivated with some care and attention in later developmental cycles. Now, who is going to give the producer a sneak preview of summative results? The staff evaluator will try, and often can do a very good job. But the role is not conducive to objectivity—not only is it dependent on the payroll (and hence one where criticism can produce resentments with which the evaluator will have to live), but it is also very quickly tied in to the production activity. Typically, the staff evaluators are the actual authors of most of the tests in curriculum products, and responsible for some of the form and content of much of the rest. Finally, the staff person is likely to have occupational tunnel-vision with respect to the effects of the materials (or methods, etc.)—that is, a tendency to look mainly in the direction of the announced goals.

Hence, it now seems to me that a producer or staff evaluator who wants good *formative* evaluation has got to use some external evaluators to get it. Using them does not render the staff evaluator redundant; on the contrary, implementation or correction of the external evaluation depends in large part on the staff person. Psychologically, the staff evaluator may find it priceless to have support from an external source for some personal—and previously unshared—worries or complaints. Now, what I have said so far supports a practice of many producers in using external evaluators. But what I have said also implies—because it springs from the hunt for objectivity/independence—the desirability of arranging goal-free conditions for the external evaluator.

As summative evaluation becomes increasingly goal-free—and I believe it will—the formative evaluation must do so to preserve the simulation. But forget that point; the same conclusion is forced on us by interest in picking up what are for the producer “side-effects.” The less the external evaluator hears about the goals of the project, the less tunnel-vision will develop, the more attention will be paid to *looking for actual effects* (rather than *checking on alleged effects*).

OTHER FAVORABLE CONSIDERATIONS

Look at the effects of considering goals on those who formulate them. It is likely to seem to them that it will pay better to err in the direction of grandiose goals rather

than modest ones—as one can see from experience in reading proposals requesting funds, where it's entirely appropriate to evaluate goals. This strategy assumes that a gallant try at Everest will be perceived more favorably than successful mounting of molehills. That may or may not be so, but it's an unnecessary noise source for the evaluator.

The alleged goals are often very different from the real goals. Why should the evaluator get into the messy job of trying to disentangle *that* knot?

The goals are often stated so vaguely as to cover both desirable and undesirable activities, by almost anyone's standards. Why try to find out what was really intended—if anything? (Similarly, the stated goals often conflict—why try to decide which one should supervene.)

A trickier point. The identification of "side-effects" with "unanticipated effects" is a mistake. Goals are only a subset of anticipated effects; they are the ones of special importance, or the ones distinctive of this project. (For example, the goals of a new math curriculum project do not usually include "employing a secretary to type up corrected copy," but of course that effect is anticipated.) Hence, "side-effects" includes more phenomena than "unanticipated effects," and some of the ones it alone includes may be important. In short, evaluation with respect to goals does not even include all the anticipated effects and gives much too limited a profile of the project. Why get into the business of trying to make distinctions like this?

Since almost all projects either fall short of their goals or over-achieve them, why waste time rating the goals; which usually *aren't* what is achieved?

GFE is unaffected by—and hence does not legislate against—the shifting of goals midway in a project. Given the amount of resentment caused by evaluation designs that require rigidity of the treatment throughout, this is an important benefit. But it's a real advantage only to the extent that the project remains within the much larger but still finite ballpark the GFE has carved out of the jungle of possible effects.

UNFAVORABLE CONSIDERATIONS— METHODOLOGICAL AND PRACTICAL

These are usually an amalgam of criticisms from various sources, sometimes real quotes.

"The GFE'r simply substitutes his own goals for those of the project." No. The GFE may use USOE's goals, or what the best evidence identifies as the needs of the nation, as standards; but simply to use his (or her) own personal preferences would obviously be to invalidate the evaluation. One needs standards of merit for an evaluation, indeed; the error is to think these have to be the goals of the evaluator or the evaluated. Another, commonly connected, error is to think that all standards of merit are arbitrary or subjective. There's nothing subjective about the claim that we need a cure for cancer more than a new brand of soap. The fact that some people have the opposite preference (if true) doesn't even weakly undermine the claim

about which of these alternatives the *nation* needs most. So the GFE may use needs and not goals, or the goals of the consumer or the funding agency. Which of these is appropriate depends on the case. But in no case is it proper to use *anyone's* goals as the standard unless they can be shown to be the appropriate ones and morally defensible.

"Great idea—but hopelessly impractical. You can never keep the evaluator from inferring the goals of the project." This is certainly false; I and others have done evaluations where only the feeblest guesses would be possible, and of no great interest. If you control the data going to the evaluator, you can obviously reduce it to the point where goals are not inferable. And interesting—not exhaustive—evaluations are still possible. An evaluator with considerable experience of goal-based evaluation does indeed find it tempting, in fact almost neurotically necessary, to reach for the security blanket of goals. But once one learns to do without it, then, like riding a bicycle or swimming without the aids one uses at first, there is a remarkable sense of freedom, of liberation.

"Why use an evaluator who only gets part of the data—you simply increase the chance that some of the most important effects (which happen to have been intended) will be missed?" Yes, this is the trade-off. The value of GFE does not lie in picking up what everyone already "knows," but in noticing something that everyone else has overlooked, or in producing a novel overall perspective. Of course, when summative time comes around, the intended effects had better be large enough to be obvious to the unaided (but expert) eye or, in general, they aren't worth very much. (The same is therefore true to a lesser extent for formative evaluation.)

"Attacking the emphasis on careful goal-formulation approaches can only lead to poor planning, a catch-as-catch-can approach, and general carelessness—which you are giving intellectual sanction." Planning and production require goals, and formulating them in testable terms is absolutely necessary for the manager as well as the internal evaluator who keeps the manager informed. That has nothing to do with the question of whether the external evaluator needs or should be given any account of the project's goals.

"I still can't see how GFE is supposed to work in practice. You can't test for all possible effects, and it's surely absurd to think you shouldn't even *bother* with testing the real goals." The external evaluator is not *there* to test goals, but rather to evaluate achievement which turns out to be conceptually distinct—and often different in practice, too. As to the idea that GFE requires testing for every possible effect, the best reply is to say that any evaluator worth hiring has to look for side-effects, and there's no limitation on where or in what form they crop up. So even the goal-based evaluator (GBE'r) has to do this allegedly impossible task. (And so, for that matter, does any applied scientist searching for the effects of a new drug—or the scientist looking for unknown causes of an important effect, e.g., death or cancer: except he searches for every possible cause, not effect.) The GFE'r looks at the treatment and/or curricular materials, after all, and can immediately formulate some hypothesis about probable effects, based on previous experience and knowledge of the research literature. Often, too, the GFE'r can look at the results of

quizzes etc., though it's desirable to do that *after* formulating the hypothesis just mentioned, to avoid premature fixation on the variables of concern to the project.

"I'm afraid the GFE is going to be seen as a threat by many producers, perhaps enough to prevent its use." It's true that even GBE was and is so threatening that its introduction has been prevented or rendered useless on many projects. But it has gradually become increasingly a requirement, and the standards for it are creeping upwards. The same is likely to be true of GFE. Now it's important to see why GFE is *more* of a threat. Primarily this is because the GFE'r is less under the control of management; not only are the main variables no longer specified by management, but they may not even *include* those that management has been advertising. The reactions by management to GFE have really brought out the extent to which evaluation has become or has come to seem a controllable item, an unhealthy situation. The idea of an evaluator who won't even *talk* to you for fear of contamination can hardly be expected to make the producer rest easy. It's probably very important, psychologically, to talk to your judge, to feel you've got across a sense of your mission, the difficulties, etc. We all have some faith in "tout comprendre c'est tout pardonner." But the evaluator isn't our judge, just the judge of something we've produced. Even if it's not much good, there's a long way to go before blame can be laid at the producer's door. If a producer really cares about quality control it won't do to insist that the project's definition of quality must be used.

METHODOLOGICAL ANALOGIES OF GFE (IN OTHER FIELDS)

The Intentional Fallacy

In the field of aesthetics it has been widely but not universally accepted that it is fallacious for a critic to consider the intentions of the artist in assessing the work of art. If the "meaning" doesn't show, it doesn't (or shouldn't) count. I am inclined to think this is a perverse view, a purist limit that goes beyond the bounds of sense. The titles of paintings, the locale of photographers, program notes at the symphony, the period of a building, even the biographies of Russian novelists, "cast new light on" the art object itself, and are interesting in themselves. The fallacy is to suppose that the only legitimate framework in which to see a work of art is as an autonomous entity. Art can enlighten, it can give pleasure, it can communicate feeling, and so on—and there's nothing in there that says the background and context of the artwork can't contribute. It's really a case where the consumer can choose. One may say that assessing the *artist* legitimately brings in these considerations, but assessing the *artwork* does not—but the slight attraction of this "tidying-up" move scarcely amounts to a compelling argument for any reasonable man.

In the educational materials production situation, on the other hand, as in the consumer field in general, we can usually establish that the intentions of the producer are of negligible concern to the consumer by comparison with satisfactory performance on the criterion dimensions (e.g., gains in reading scores). Not only is

this so, but there seems to be little reason why it *shouldn't* be so. When the history of educational R&D is written (if ever historians can be found to stoop to such a low-status task which happens to be socially valuable) then the intentions of producers will be of great interest. For the future producer, a study of these may be far more valuable than a study of the products.

So the “intentional fallacy” is not, in my view, a fallacy in the area where the term was introduced—but it would be one in the evaluation of consumer goods.

Motives and Morality

A tremendous tension has long existed in philosophical ethics between those who believe that the morality of acts is principally determined by their motivation (“He meant well”) and those who would assess acts in terms of their consequences alone (“Write *that* on his gravestone; first, he should be shot”). Current pop ethics is on the conscience trip—the “pragmatist” is seen as the opposition.

The special feature of this case is that the act involves the motive in a much more intimate way than the product involves the producer’s intent. It has been argued that the same physical motions performed with different intentions are definitionally a different act; the distinctions between manslaughter and murder, between borrowing and theft, erring and lying, for example, are said to be distinctions between different acts. One cannot argue that a programmed text supposed to teach economics better than the competition but which actually teaches reading better (and economics the same) is crucially different for the consumer from one in which the side-effect was the primary aim of the producer. And it is for just this reason I prefer the role of the GFE’r for summative evaluation.

On the philosophical issue: I prefer to say that neither exclusive position is defensible, that the issue is resolved one way or the other in *particular* cases where the *point* of the evaluation becomes clear.

Double-Blind Designs

A correspondent writes, “The so-called ‘double-blind’ medical experiment isn’t blind in terms of goal or purpose. A treatment is being tested for its effect on a specific disease. The ‘blind’ is strictly in terms of the S’s or E’s knowledge of who is getting what treatment. Thus I think your use of the analogy is inappropriate.” The analogy is not intended to be an identity. The point of the analogy is to remind one that medical research, until the scurvy study, ignored the error due to the agent and evaluator knowing that the treatment being given to a particular patient was a dummy. Not only did this affect the agent’s behavior in giving it, but it affected the evaluator’s care in assessing the effects. After all, how could one seriously look for therapeutic results from a sugar-pill? “Blinding” the assessor made the search equally careful in both cases. Analogously, “blinding” the educational evaluator

ensures (to the maximum possible extent?) equal care in looking for effects that happen not to have been goals. Now it's true that the GFE'r may make it the first order of business to infer the goals of the producer. In fact, that's what happened in the second GFE study of which I have received details. (But in the medical case this is often possible, too. In 1958 or so I spent a great deal of time refining placebo effect research designs; the problems of matching for the taste and side-effects of the experimental drug, amongst other difficulties, are typically not solvable.) All one can do is to make it as hard as possible. In particular, one can try to cut out cues which allow inference of intent other than via noticing success. It's not disastrous if the medical researcher infers *from the results* that treatment B must have been the new medication, treatment A the placebo. The inference may or may not be correct; it can only be damaging if it is made *during* the experiment and hence might influence the later procedures. But even that possibility can usually be handled by splitting the role of recorder from that of agent. By analogy, we cannot get too worried about an evaluator who, seeing massive gain scores on an addition-of-integers test, infers that a major goal of the materials was to improve addition of integers. On the other hand, we must try to avoid having the evaluator come to this conclusion by reading the introduction to the materials, because that is likely to corrupt his later perceptions. When the evaluator devises special instruments for assessing inventory on a parameter that has not previously been tested, we can isolate the role of the agent doing the testing from the role of the scorer, and we can arrange that the scorer does not know the pretests from the posttests, or the experimental group's tests from the control group's tests.

In the early GFE just mentioned, where the evaluator worked diligently to reconstruct the goals, he was doing this by observing various effects which seemed desirable. He concluded that these were probably intended. But the step of inferring goals was totally unnecessary—he could just as well have left the matter by noting the desirable results. Similarly, where he inferred failure (e.g., at teaching the inquiry approach) he could just as well have made no comment, or noted lack of performance in this desirable dimension, from which the evaluend can conclude failure.

Finally, although it is typical of the medical situation that a major parameter is identified in advance, no evaluation of drugs today can avoid the search for side effects, from the most remote area of the symptom-spectrum. Nor is this obligation restricted to Federal checks; the formative evaluation of drugs requires that the manufacturer run studies that are both double-blind and side-effect sensitive. It would not be difficult to run these evaluations goal-free, but it has little point; given only the characteristics of the patients to be treated, the goal of the treatment would be fairly obvious. In education, the situation is different—more like preventive medicine.

In sum, I think there's an illuminating analogy between the move to double-blind methodology and the (further) move to GFE. The gains from double-blind were not significant in the physical sciences—it was an innovation of great value to medicine—but it is an innovation that may pay off in education.